LA-UR-21-22528

Approved for public release; distribution is unlimited.

Title:	Pilot 1 Uncertainty Quantification: A critical appraisal
Author(s):	Anghel, Marian
Intended for:	Report
Issued:	2021-03-15

Pilot 1 Uncertainty Quantification
A critical appraisal
Feb 2021

Marian Anghel

LANL

- **Goal:** Obtain estimation of uncertainty, i.e. a confidence interval for predictions made by the model, specifically for neural network regression.
- **UQ Models:**
  - Homoscedastic: - Gaussian noise model, no feature dependence (homogeneous noise)
    - We only learn $y = f(x)$ and estimate a uniform uncertainty σ

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \|y_i - f(\mathbf{x}_i)\|^2 \qquad \sigma^2 = \frac{1}{M}\Sigma_{j=1}^{M}(y_j - f(x_j))^2$$

  - Heteroscedastic: - Gaussian noise model, smooth feature dependence (heterogeneous noise)
    - We learn $y = f(x)$ and $\sigma(x)$.

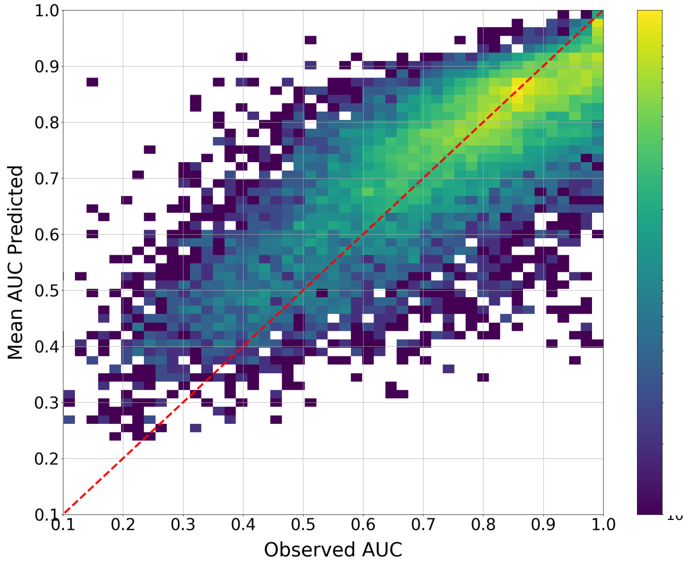
$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2\sigma(\mathbf{x}_i)^2} \|y_i - f(\mathbf{x}_i)\|^2 + \frac{1}{2}\log\sigma(\mathbf{x}_i)^2$$

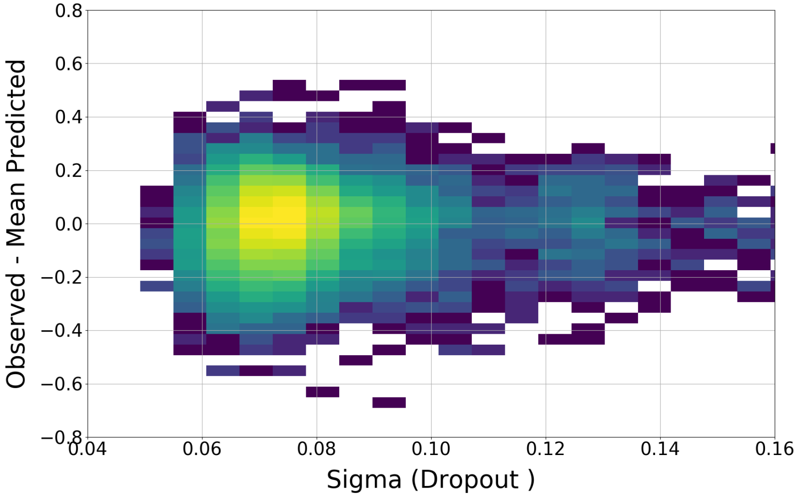
  - Quantile: - No explicit noise model, assumes smooth quantile functions
    - We learn $f_\alpha(x)$ for $\alpha = 0.1, 0.5, 0.9$ (the 1st, 5th, and 9th deciles) with $y = f_{0.5}(x)$

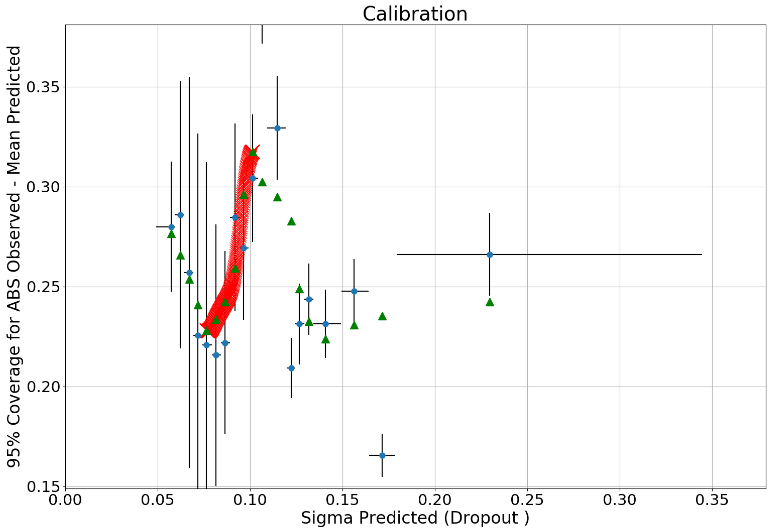
$$\mathcal{L}(y,f) = \sum_{\alpha} \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_\alpha\left(y_i - f_\alpha(\mathbf{x}_i)\right) \qquad \mathcal{L}_\alpha(\xi_i) = \begin{cases} \alpha\, \xi_i & \text{if } \xi_i \geq 0\,, \\ (\alpha - 1)\, \xi_i & \text{if } \xi_i < 0\,. \end{cases} \qquad \xi_i = y_i - f_\alpha(\mathbf{x}_i)$$

- **Lessons learned:** A retrospective analysis of UQ estimates for drug response models.
  - Calibration curves are beset with high uncertainty and do not provide reliable, local, information about uncertainty.

GDSC Data – Heteroscedastic - 10% dropout



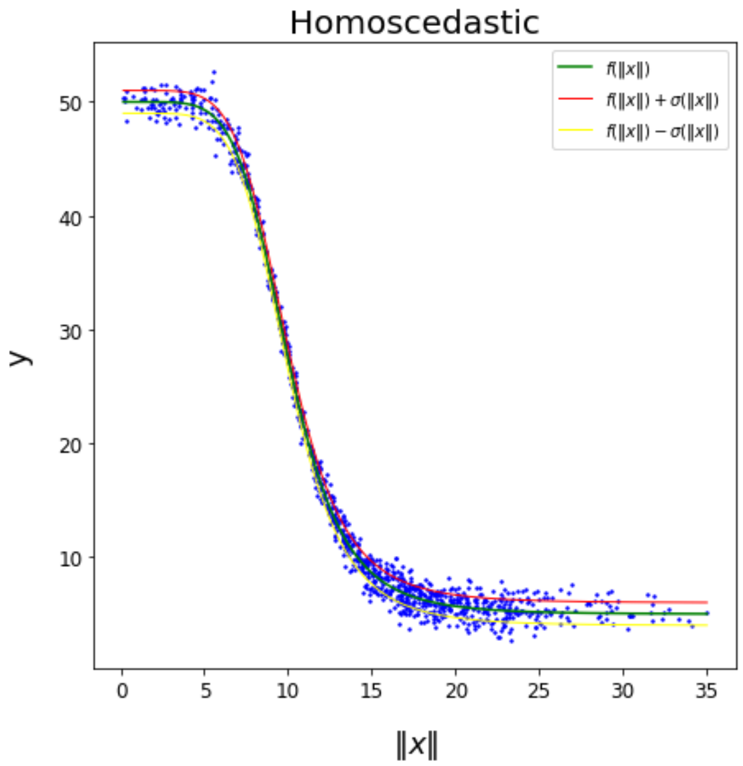


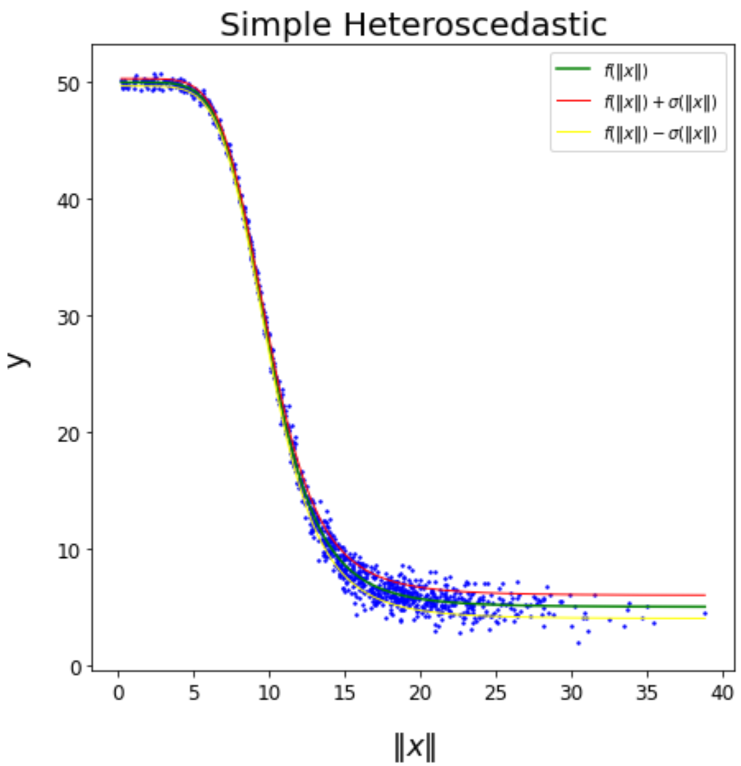


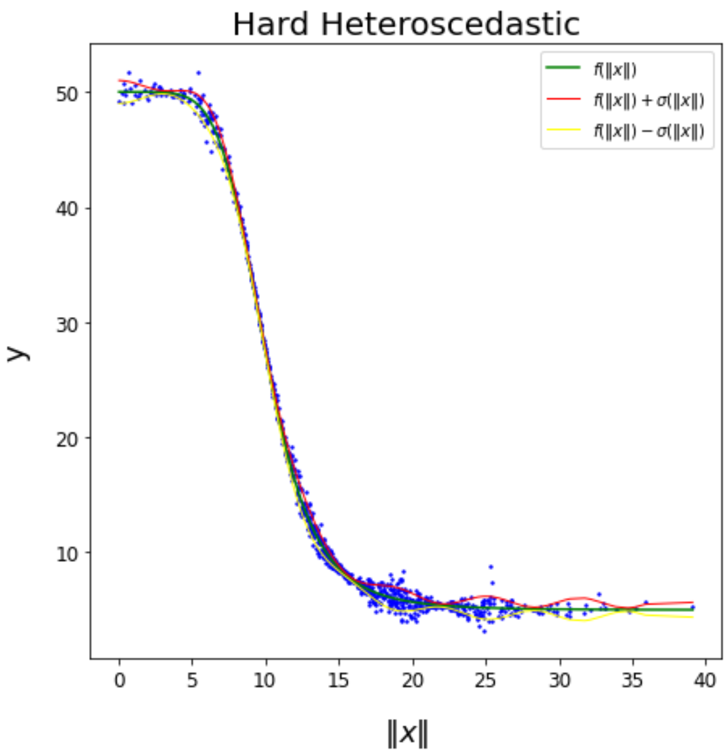
Blue: Computed empirical relationship between predicted STD and error.
Green: Smoothing of empirical relationship.
Red: Range of monotonically increasing relationship.

- The trend in the data can be used for calibration but it has very large uncertainty and cannot provide reliable local information for prediction and experiment design.
- The uncertainty of the uncertainty estimation is very large!

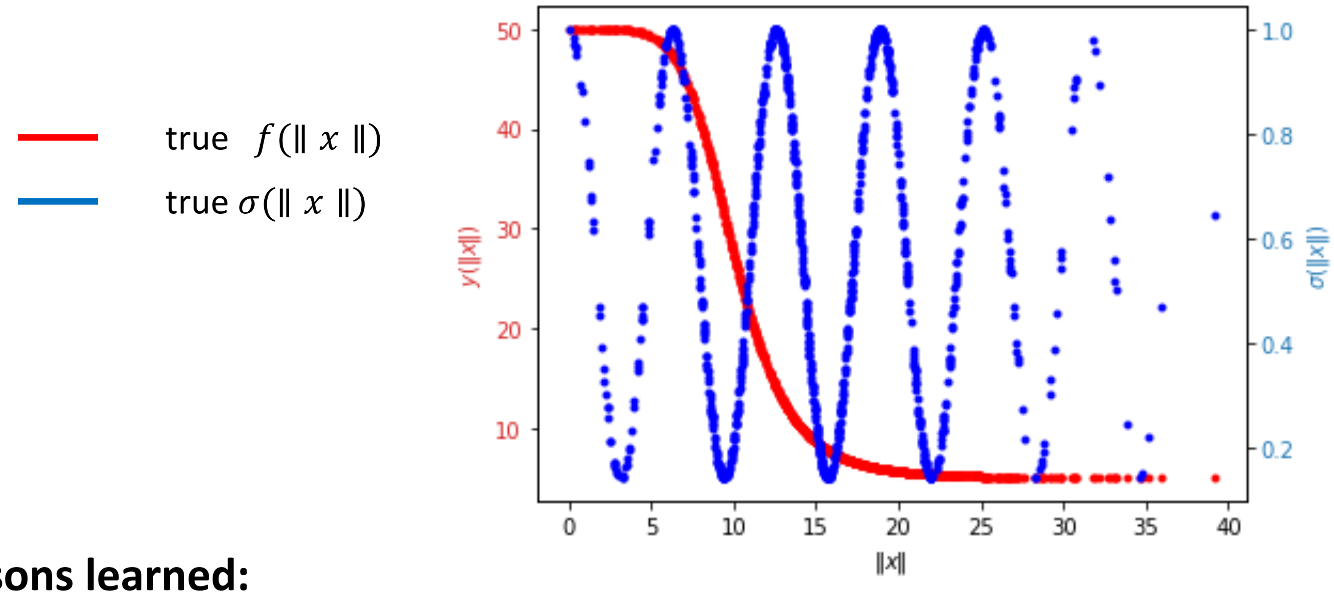
- **Lessons learned:** We designed and analyzed synthetic data sets for which we know both the response $y = f(x)$ and the noise $\sigma(x)$.
  - This is the only way in which any lessons could be learned.

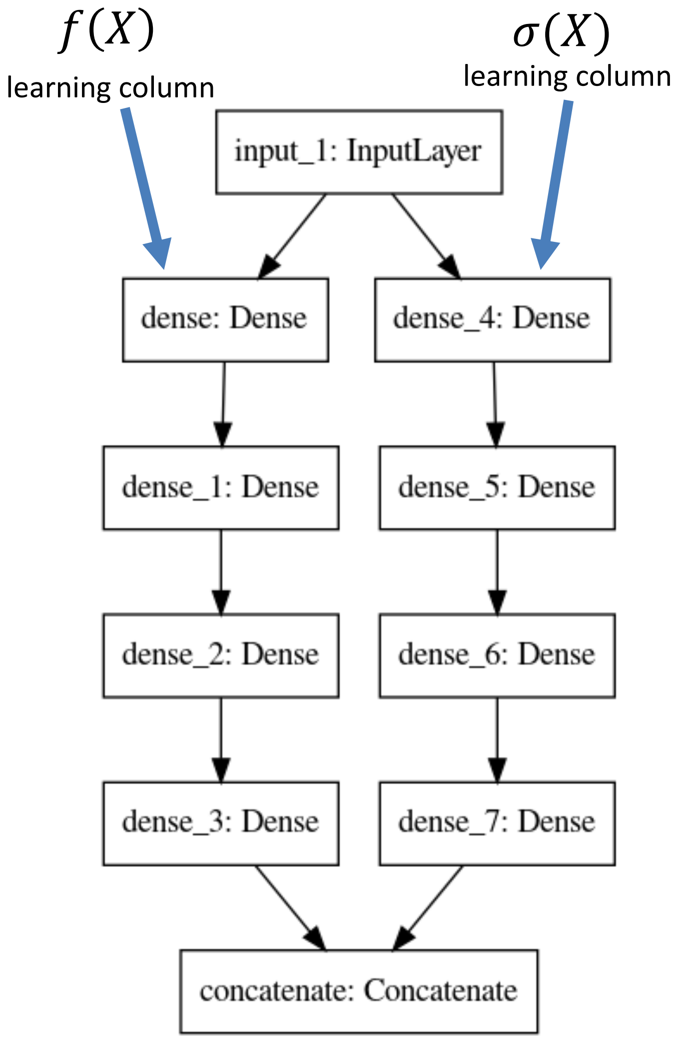






- By design the response and uncertainty depend on the norm $\| x \| = \sqrt{\sum_{i=1}^{D} x_i^2}$ of the feature vector $x \in \mathbb{R}^D$, so that we can visualize the performance of the learning algorithms.
- Conceivably, this may bias our conclusions.



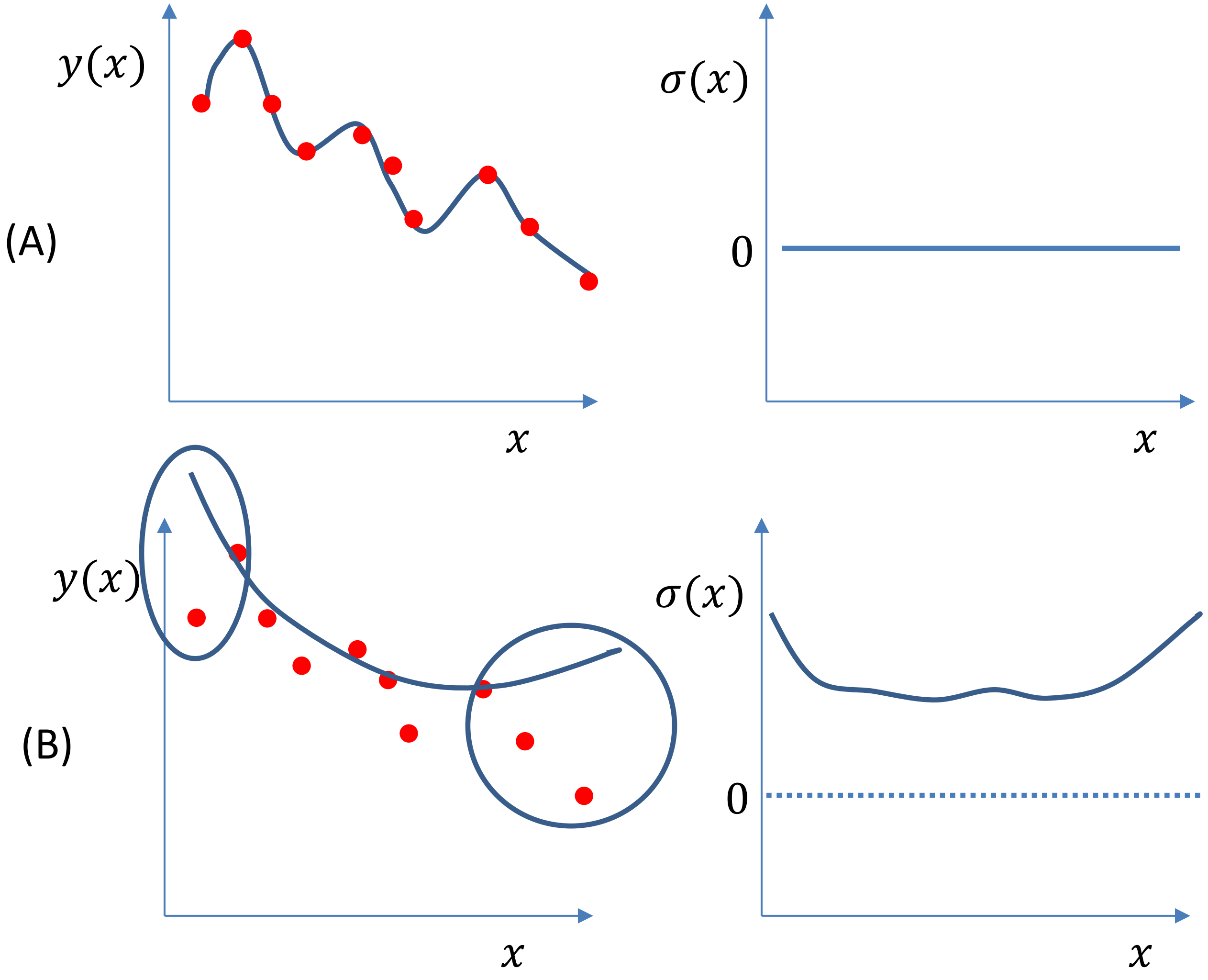


- **Lessons learned:**
  - The response and the uncertainty can have different learning complexities (the number of training-samples that we need in order to successfully learn each target function).
  - Heteroscedastic and quantile formulation with independent, or loosely coupled, prediction of mean and variance to allow for different regularization functions or even architecture for mean and variance outputs.
  - We should implement a model that enables us to explore the tradeoff between these two complexities.
  - This complicates the selection of hyperparameters.
  - This complicates learning itself, since the learning rate of each target function is different.

The loss function couples two learning tasks

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2\sigma(\mathbf{x}_i)^2} \|y_i - f(\mathbf{x}_i)\|^2 + \frac{1}{2}\log \sigma(\mathbf{x}_i)^2$$

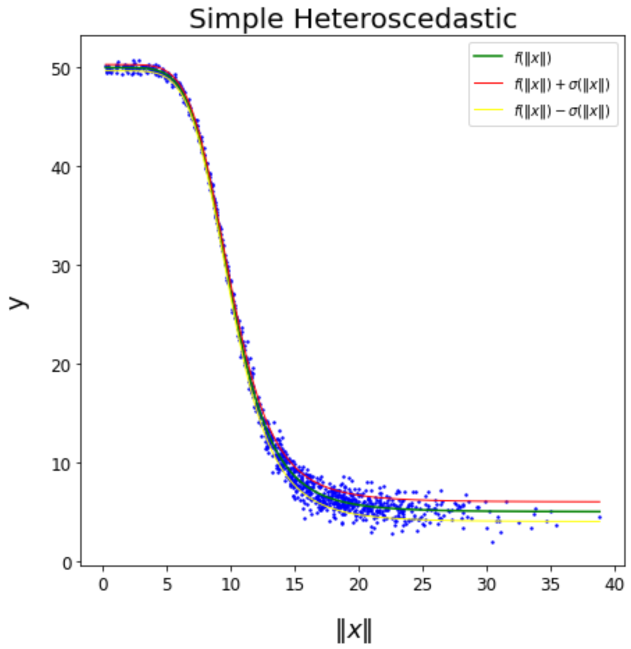
- **Lessons learned:**
  - Model bias is confused with variance (uncertainty): Very difficult to disentangle the two when we have only one measurement per observation point.



- (A) Model is too complex and overfitting the data.
  The estimated $\sigma(x)$ is very small.
- (B) Model is too simple and biased.
  The estimated $\sigma(x)$ is large to accommodate bias.
- Cross-validation methods should help select the right model complexity: Is this truly the case?
- Often, the bias is “local” and confuses the validation error estimate and the correct selection of model complexity.

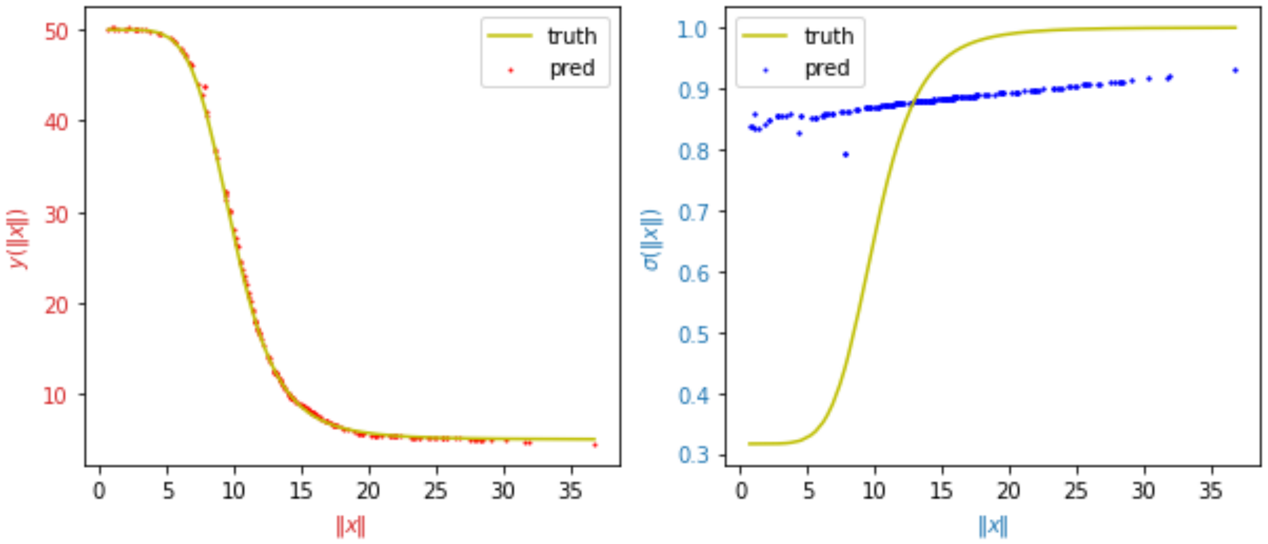
- **Lessons learned:**
  - While all models can learn an accurate estimation of response, the accurate estimation of uncertainty is very difficult.
  - Model bias is confused with variance (uncertainty): Very difficult to disentangle the two when we have only one measurement per observation point.

Training data



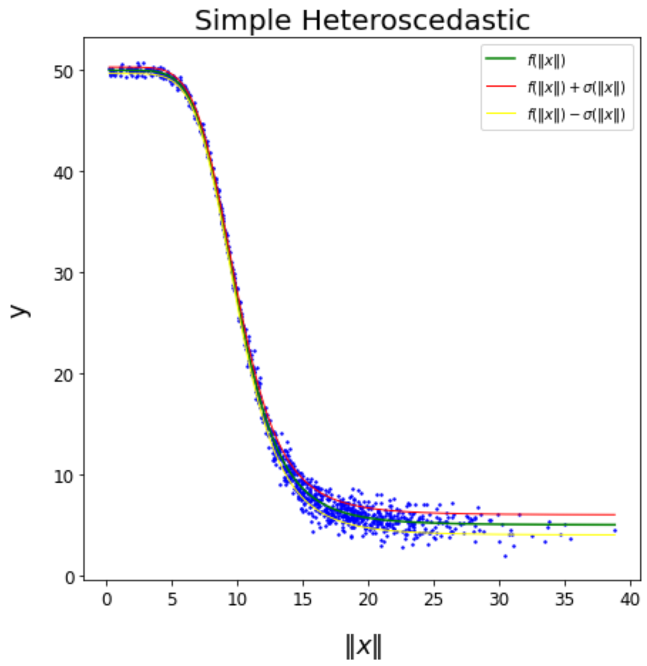
- Simple heteroscedastic data
- dim X = 1 (feature space dimension)
- one, noisy measurement everywhere

Modeling performance

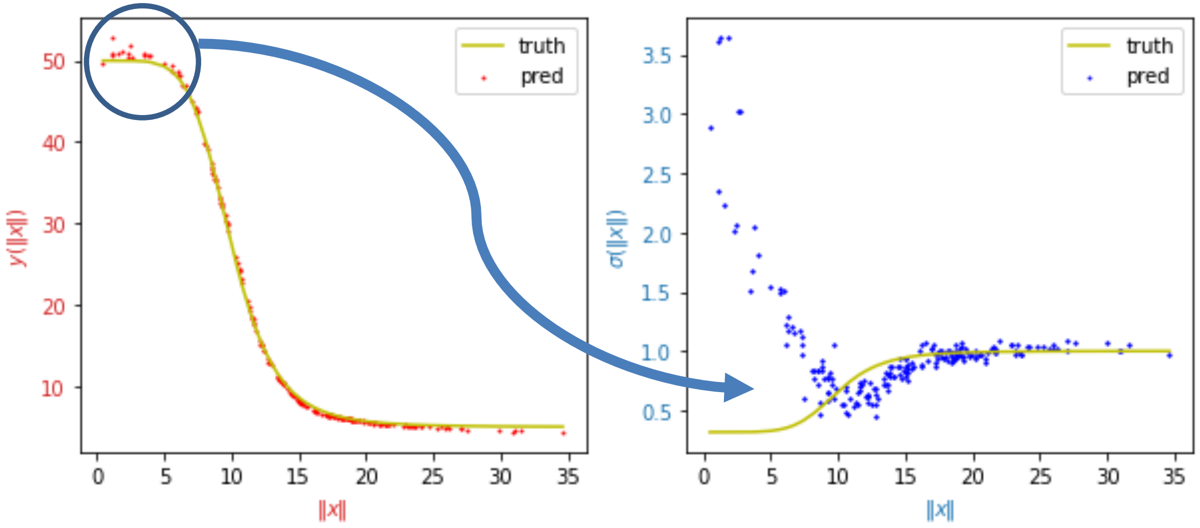


- Heterogeneous learning model
- Learns well the response
- The uncertainty estimated is flat.

- **Lessons learned:**
  - The quantile model has best performance for estimating uncertainty.
  - Model bias is confused with variance (uncertainty): Very difficult to disentangle the two when we have only one measurement per observation point.



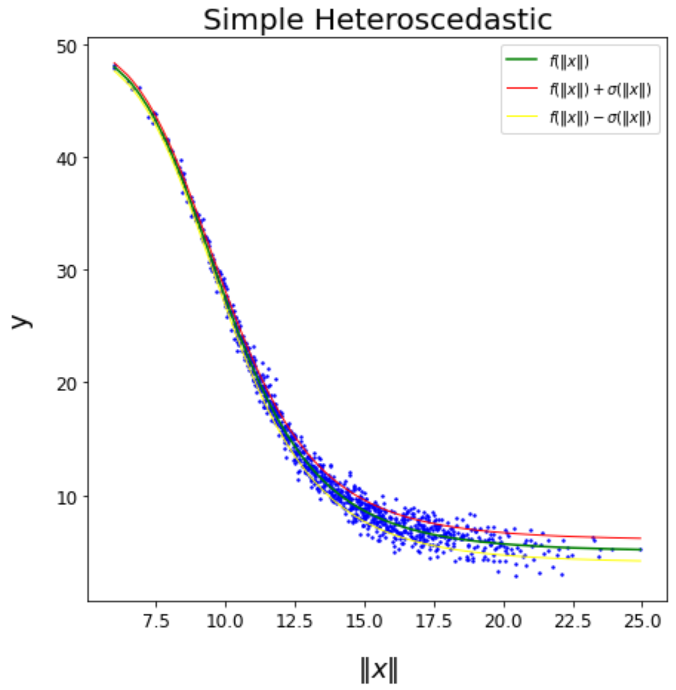
- Simple heteroscedastic data
- dim X = 1 (feature space dimension)
- one, noisy measurement everywhere



- Quantile learning model.
- Learns well the response.
- The uncertainty estimated is very good over a subset of the feature space.
- Subtle trade-off between bias (at small X values) and variance (noise).
- This trade-off is often local and hard to detect.

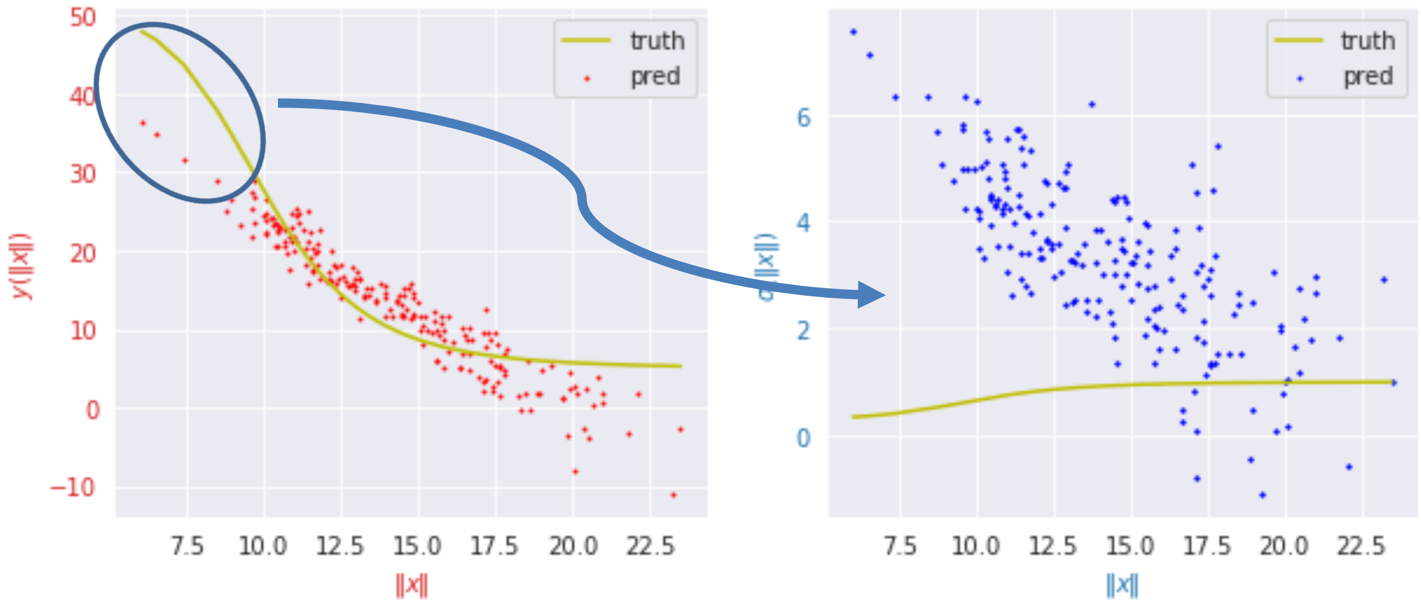
- **Lessons learned:**
  - Learning response curves well is harder in high dimensional feature spaces, but learning uncertainty is even harder.

Training data



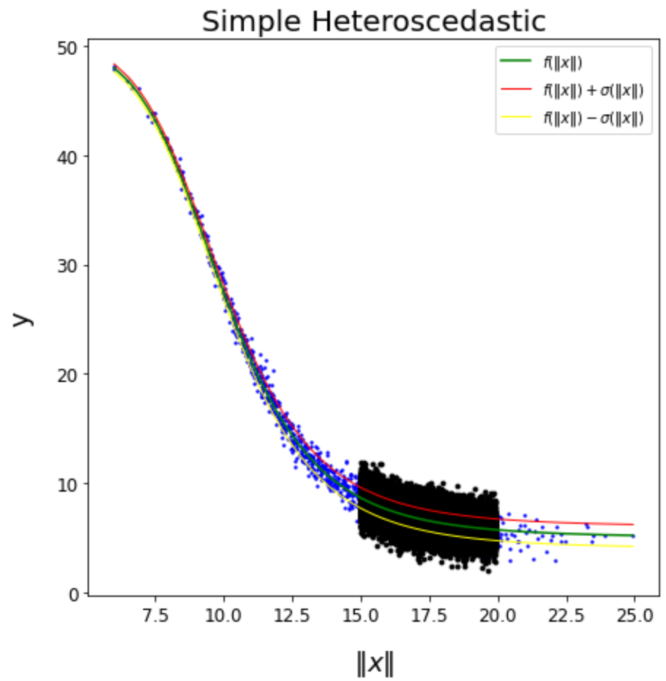
- Simple heteroscedastic data
- dim X = 10 (feature space dimension)
- one, noisy measurement everywhere

Modeling performance

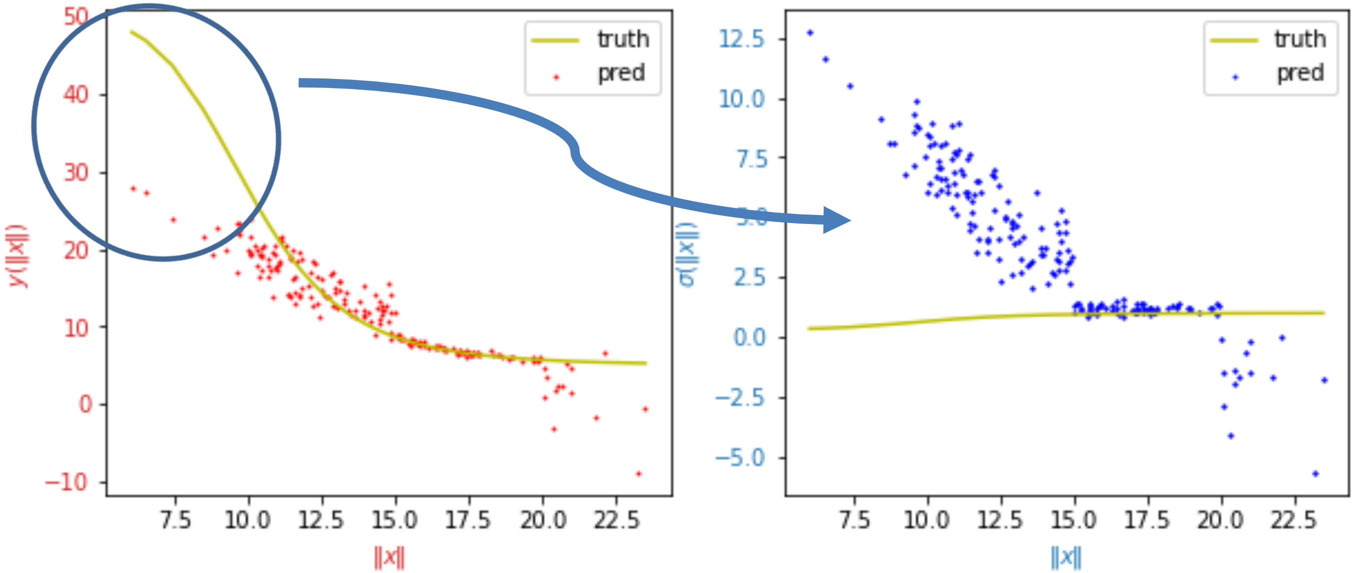


- Quantile learning model.
- Both response and uncertainty are hard to estimated.
- Model bias at small X values is confused for noise.

- **Lessons learned:**
  - Introducing multiple measurements in a limited domain of the feature space (100 replicates for each training point with $\| X \| \in [15,20]$ improves estimation at a high experimental cost.
  - There is a negative impact on response estimation elsewhere.
  - The experimental cost for learning uncertainty is very large.



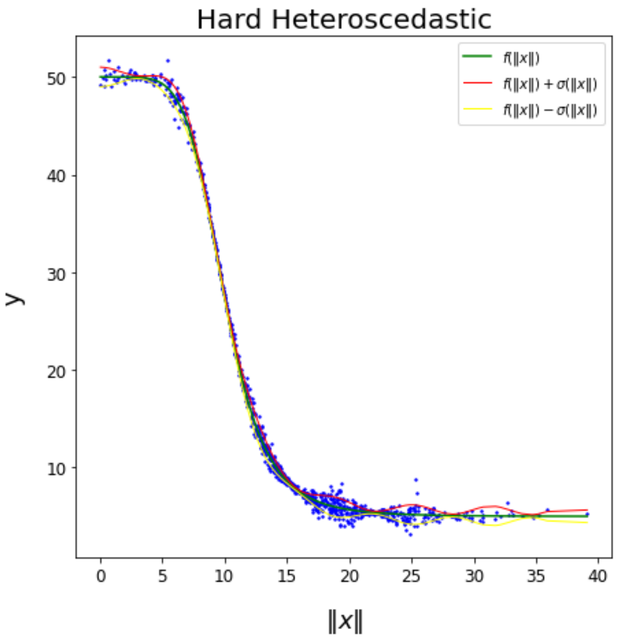
- Simple heteroscedastic data
- dim X = 10 (feature space dimension)
- 100 noisy measurements for for each training point with $\| X \| \in [15,20]$



- Quantile learning model.
- Improved response and uncertainty estimates in region where multiple (100) replicates are introduced.
- Model bias at small X values is confused for noise.
- Negative $\sigma(X)$ for large X values due to quantile inversion, i.e.

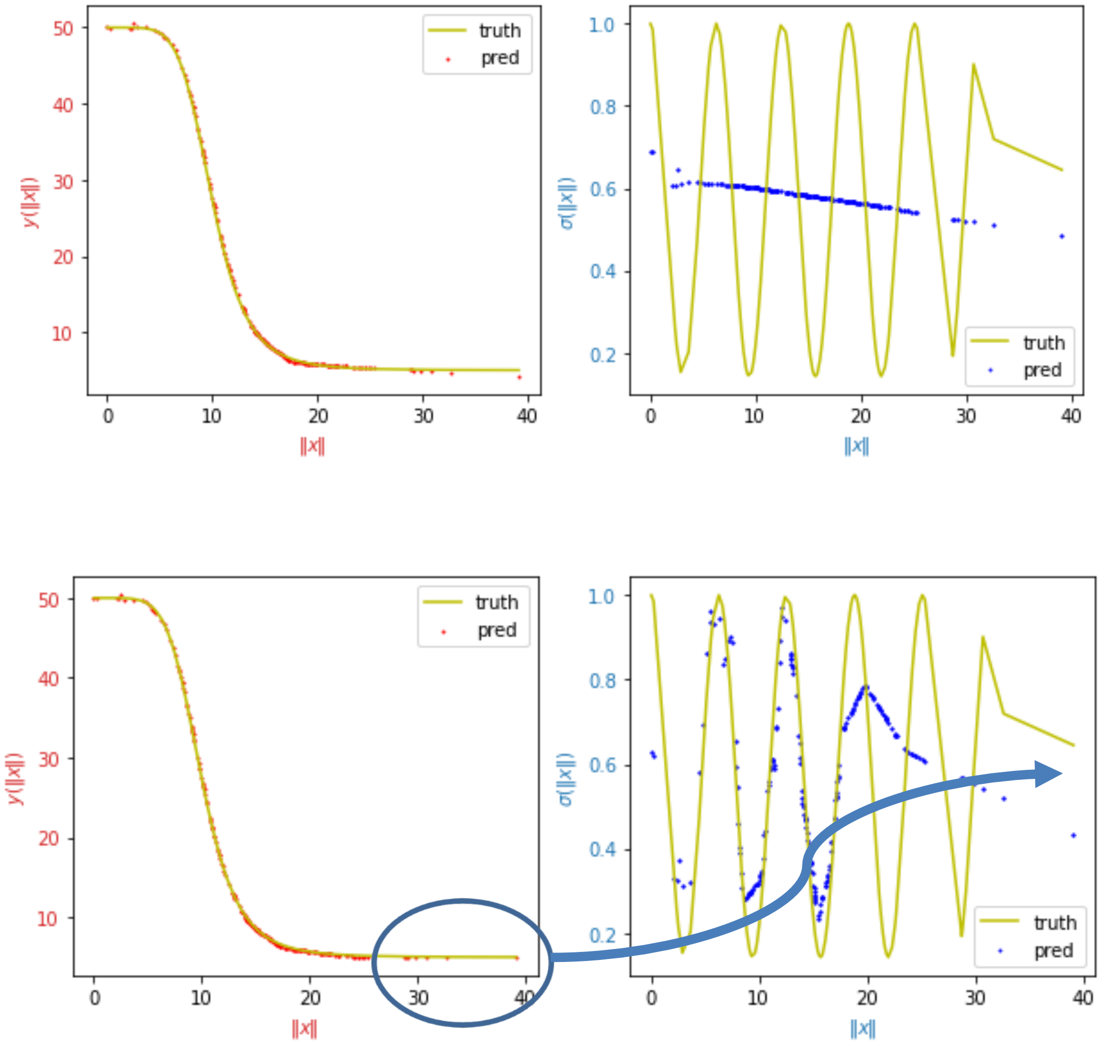
$$f_{0.9}(x) < f_{0.1}(x)$$

Training data



- Hard heteroscedastic data
- dim X = 1 (feature space dimension)
- one, noisy measurement everywhere

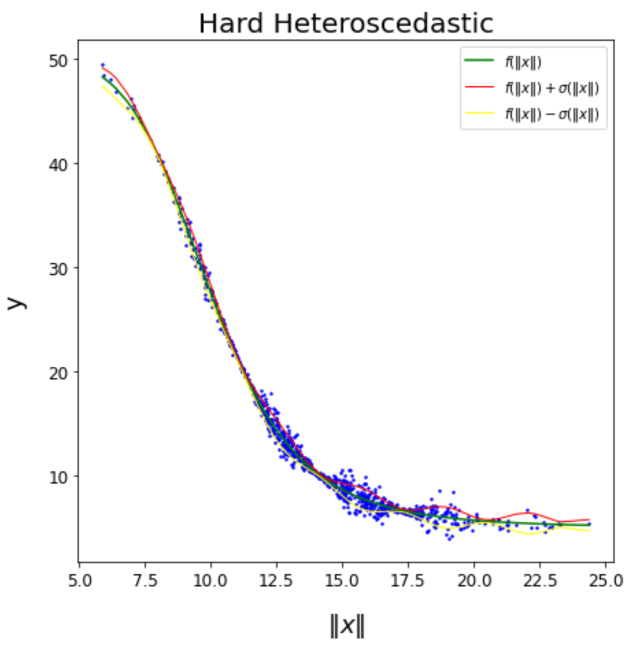
Modeling performance



- Heteroscedastic learning model
- Flat uncertainty estimation
- Captures the correct scale of the uncertainty

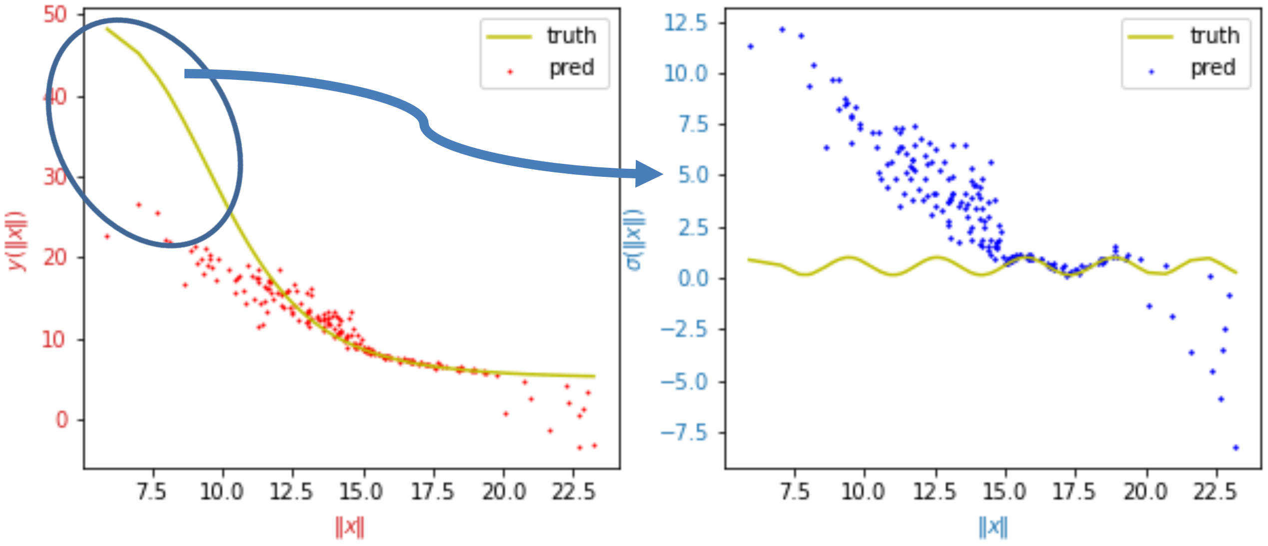
- Quantile learning model.
- Learns well the uncertainty.
- Large error for large ‖ X ‖ values due to scarce training data

Training data



Hard heteroscedastic data
- dim X = 10 (feature space dimension)
- one, noisy measurement everywhere

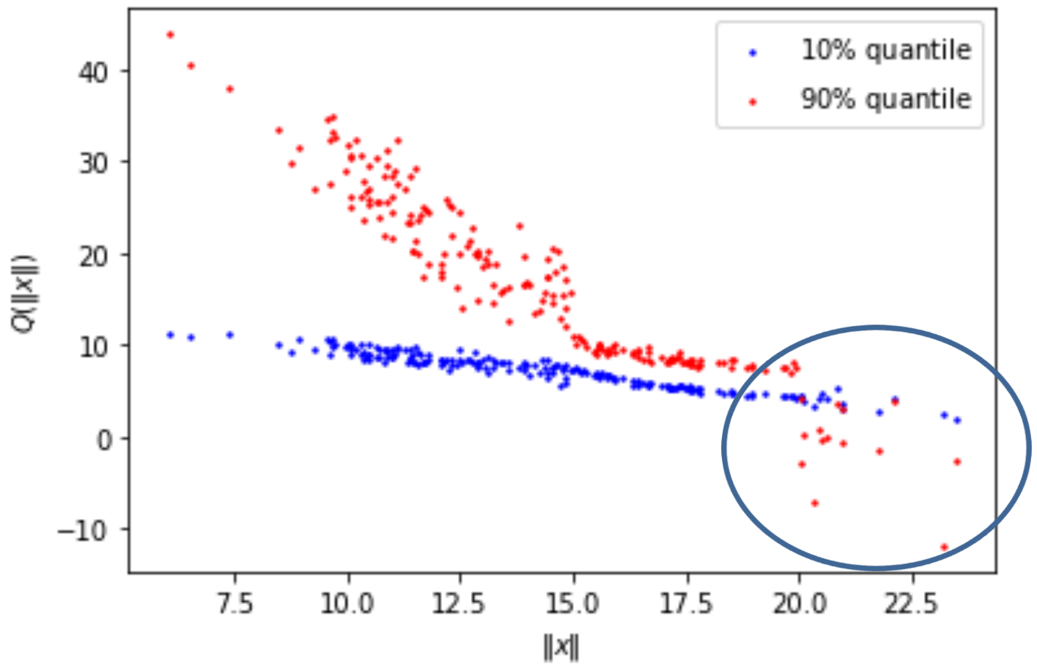
Modeling performance



- Quantile learning model.
- Better response and uncertainty estimates in region with higher density of training data.
- Model bias at small $\| X \|$ values is confused for noise.
- Negative $\sigma(X)$ for large $\| X \|$ values due to quantile inversion, i.e.

$$f_{0.9}(x) < f_{0.1}(x)$$

- **Lessons learned:**
  - The current implementation of the models allows the response and sigma learning function spaces to have different complexities (reg_l2 params). Thus we explore the tradeoff between these two complexities: Can we detect bias?
  - The quantile model exhibits quantile inversion in regions of the feature space with sparse training examples. This can be used to detect regions in feature space where more measurements are needed.
  - Alternatively, we should include ordering constraints in the loss function to prevent this from happening.



For large $\| X \|$ values we notice a quantile inversion effect, i.e. $f_{0.9}(x) < f_{0.1}(x)$, flags the region of *uncertain* uncertainty.

- **Lessons learned:**
  - It is very difficult to optimize these networks to learn uncertainty, while good approximations to the response curve can be found for a large range of network (hyper)parameters.
  - The model learns the response first (since there is direct information used for learning) and stops improving the learning of uncertainty (for which learning information is not usually available).

- **New learning strategies:**
  - We need improved learning strategies: For example, in a model with distinct learning columns for $f(X)$ and $\sigma(X)$ we can alternate between learning $f(X)$ while weights for learning $\sigma(X)$ are frozen, and vice versa. What is a good learning strategy?

- **Concluding remarks:**
  - Accurate estimation of uncertainty remains a difficult, open problem.
  - The trade-off between bias and variance is often local and difficult to detect.
  - The experimental cost for learning uncertainty is very large.
  - Improved uncertainty estimation impacts the learning of a good response model, and vice versa. The two learning objectives may be competing against each other.
  - With so much data needed to learn well uncertainty, we may just learn uncertainty as an independent learning problem.
  - How to approach uncertainty estimation when this is not the case?